GUEST EDITORS' INTRO

SEARCHING THE INTERNET

Robert E. Filman
Lockheed-Martin Corporation
Sangam Pant
Lycos, Inc.

Networks and devices are going to get faster and cheaper. What will continue to be AI-hard is the problem of making sense of the mass of data and misinformation that fills the Web.

Ten years ago, it was straightforward to predict the direction of computer technology. The hardware would keep getting faster and cheaper. Similarly, software development would continue to push the frontier of the feasible—interfaces would get more immersive, databases bigger, simulations more fine grained. The surprise, from a technologist's point of view, has been the social phenomena of the World Wide Web. A decade ago, the general population was computer phobic and took their entertainment passively. Now we see advertisements trumpeting HTTP addresses and grandmothers surfing the Net. Meanwhile, everyone and their cousin has turned into a publisher of original HTML material.

The networks and devices are going to continue to get faster and cheaper. Our current concerns with network protocols and algorithms for, say, real-time multimedia presentations will seem in 30 years as quaint as tape-sorting algorithms do today. What will continue to be AI-hard is the problem of making sense of the swelling mass of data and misinformation that fills the Web.

That's the problem of searching the Web.

DIRECTORIES AND SPIDERS

In the meantime, Web search is as big an industry as any on the Internet. The search industry has evolved two dominant ways to find things: directories and spiders.

In the spirit of Melvil Dewey, directories order all knowledge into some structure and classify individual Web pages with respect to that structure. Prominent commercial directories are Yahoo and Magellan. The problems with directories are that (1) classification is a labor-intensive activity, and there are far more publishers on the Web than classifiers, and (2) if the information you seek is not reflected by the classification structure, you're out of luck.

The alternative is intensive automation, where computer programs can (ultimately) look at everything and organize it every which way. Commercially, these are seen in sites such as AltaVista, WebCrawler, Excite, Infoseek, Lycos, and HotBot. Such systems have three important parts:

- a "robot" or "spider" that explores the Web and finds pages;
- a database of information refined about those pages, including the ability to test that database against queries and order the resulting matches; and
- a user interface for obtaining queries and presenting the results.

In many respects, the critical technology is the second of these: organizing and searching (primarily) ill-structured textual information. The recent prominence of the Web not withstanding, text search engines have been around for a long time, centered in the field of Information Retrieval. From IR we get the concepts of recall (the percentage of all relevant documents retrieved) and precision (the percentage of the documents retrieved that are actually relevant). (See this issue's Arachnoid Tourist for an illustration of the relative precision and recall of the current commercial favorites.)

The alternative to classification is intensive automation, where computer programs locate and organize everything.

IR also gives us algorithms for search with respect to Boolean combinations of these specific properties of texts and with respect to statistical measures of texts as a whole. The literature is filled with classification algorithms; *IEEE Internet Computing* published a good overview of the possibilities last Fall in Gudivada et al.[1]

Commercial search engines have embraced IR technology. The major difficulties have been adapting techniques that were honed on well-specified domains like medicine or law to the amorphous, all-encompassing Web with the orders of magnitude larger databases, a flood of requests, and the difficulties of obtaining distributed information. The Web is estimated to be anywhere from 400-500 million documents large, and the major commercial search engines receive 15-20 thousand queries per minute. Companies offering search engines on the Internet find that in the presence of a multitude of matches, most Web surfers check only the first few matches for any query. The algorithms that order the matching documents for queries are usually among a search firm's most closely guarded trade secrets.

SEARCHING THE SEARCH ENGINES

How can we respond, technologically, to these issues? One way is through coalescing the strengths of different search engines by programming "metasearch" engines that add a layer of analysis and synthesis between the user and a set of databases.

Two of the articles in this issue present such solutions. Lawrence and Giles in "Context and Page Analysis for Improved Web Search" (pp. 38-46) describe the NECI search engine, which examines the actual pages suggested by multiple underlying search engines and presents the results with respect to the user's actual query.

Benitez et al., in "Using Relevance Feedback in Content-Based Image Metasearch" (pp. 59-69), introduce a metasearch engine that refines its presentations according to the past performance of its constituent engines, thereby learning what works and how to better handle queries for particular users. They illustrate this work with respect to retrieving images, an area with additional complexities all its own.

BEYOND TEXT

In contrast to classical IR, Web documents are not just text. Internet search engines can take advantage of the structure of the Web itself. Li's article, "Toward a Qualitative Search Engine" (pp. 24-29), illustrates one such approach: searching the text of "what points to a page." Over time, documents referred to by many other sites get higher weights; the hyperlink text adds context to what otherwise might be just a popularity contest with the whole world as a critic.

Of course, unstructured publication is an accident of technological evolution, not an imperative. Free unstructured text may give way to more structured documents as standards like XML take off and electronic commerce becomes a force. Structured text will make the job of precision easier, but it opens up the can of worms of universal schema definitions. Lassila's article, "Web Metadata: A Matter of Semantics" (pp. 30-37), explores the concepts of giving documents more structure through "meta" information. Specifically, Lassila describes the Resource Definition Framework (RDF), an architecture for supporting metadata on the Internet and WWW in development with the World Wide Web Consortium. Lassila is editor of the W3C working draft on RDF.

E-COMMERCE

Much of the Web is about money, and the search for ways to efficiently link products and their properties is a major opportunity to advance the cause of search technologies on the Web. In "Case-Based Reasoning Support for Online Catalog Sales" (pp. 47-54), Vollrath et al. demonstrate how the mechanisms of case-based reasoning can be applied to finding better fits between available products and customer requirements.

And finally, Norvig's "Virtual Database Technology" (pp. 55-58) applies the metasearch paradigm to a commercial product for doing comparative price shopping across the Web.

THE FUTURE OF THE FIELD

As the articles in this issue illustrate, Internet search spans research in information retrieval, machine learning, database systems, user interfaces, and artificial intelligence. Advances in all these disciplines will affect the development of search technology. We hope you enjoy this issue and use it to glimpse the future of this field. ■

REFERENCE

1. V.N. Gudivada et al., "Information Retrieval on the World Wide Web," *IEEE Internet Computing*, Vol. 1, No. 5, Sept. 1997, pp 58-68; available to Computer Society Digital Library subscribers at http://computer.org/internet/.

Robert E. Filman is a staff software engineer in the Advanced Technology Center of Lockheed-Martin in Sunnyvale, California, currently working with Microelectronics and Computer Technology Corporation (MCC) to develop the next generation of distributed application infrastructure. He received a PhD in computer science from Stanford University. He is a member of the IEEE Internet Computing editorial board.

Sangam Pant is vice president of engineering for Lycos, Inc. He received a BE in electronics engineering from Maharaja Sayajirao University in Baroda, India. Pant has received three patents and authored numerous papers in the field of database systems and distributed computing..